DORO: Distributional and Outlier Robust Optimization

Runtian Zhai [* 1]   **Chen Dan** [* 1]   **J. Zico Kolter** [1]   **Pradeep Ravikumar** [1]

Abstract

Many machine learning tasks involve subpopulation shift where the testing data distribution is a subpopulation of the training distribution. For such settings, a line of recent work has proposed the use of a variant of empirical risk minimization(ERM) known as distributionally robust optimization (DRO). In this work, we apply DRO to real, large-scale tasks with subpopulation shift, and observe that DRO performs relatively poorly, and moreover has severe instability. We identify one direct cause of this phenomenon: sensitivity of DRO to outliers in the datasets. To resolve this issue, we propose the framework of DORO, for Distributional and Outlier Robust Optimization. At the core of this approach is a refined risk function which prevents DRO from overfitting to potential outliers. We instantiate DORO for the Cressie-Read family of Rényi divergence, and delve into two specific instances of this family: CVaR and χ^2-DRO. We theoretically prove the effectiveness of the proposed method, and empirically show that DORO improves the performance and stability of DRO with experiments on large modern datasets, thereby positively addressing the open question raised by (Hashimoto et al., 2018). Codes are available at https://github.com/RuntianZ/doro.

1. Introduction

Many machine learning tasks require models to perform well under distributional shift, where the training and the testing data distributions are different. One type of distributional shift that arouses great research interest is *subpopulation shift*, where the testing distribution is a specific or the worst-case subpopulation of the training distribution. A wide range of tasks can be modeled as subpopulation shift problems, such as learning for algorithmic fairness (Dwork et al., 2012; Barocas & Selbst, 2016) where we want to test model's performance on key demographic subpopulations, and learning with class imbalance (Japkowicz, 2000; Galar et al., 2011) where we train a classifier on an imbalanced dataset with some minority classes having much fewer samples than the others, and we want to maximize the classifier's accuracy on the minority classes instead of its overall average accuracy.

Distributionally robust optimization (DRO) (Namkoong & Duchi, 2016; Duchi & Namkoong, 2018) refers to a family of learning algorithms that minimize the model's loss over the worst-case distribution in a neighborhood of the observed training distribution. Generally speaking, DRO trains the model on the worst-off subpopulation, and when the subpopulation membership is unknown, it focuses on the worst-off training instances, that is, the tail performance of the model. Previous work has shown effectiveness of DRO in subpopulation shift settings, such as algorithmic fairness (Hashimoto et al., 2018) and class imbalance (Xu et al., 2020).

However, in our empirical investigations, when we apply DRO to real tasks on modern datasets, we observe that DRO suffers from poor performance and severe instability during training. The issue that DRO is sensitive to outliers has been raised by several previous papers (Hashimoto et al., 2018; Hu et al., 2018; Zhu et al., 2020). In this paper, we study the cause of these problems with DRO, and develop approaches to address them.

In particular, we identify and study one key factor that we find directly leads to DRO's sub-optimal behavior: DRO's sensitivity to outliers that widely exist in modern datasets. In general, DRO maximizes a model's tail performance by putting more weights on the "harder" instances, i.e. those which incur higher losses during training. On the one hand, this allows DRO to focus its attention on worst-off subpopulations. But on the other hand, since outliers are intuitively "hard" instances that incur higher losses than inliers, DRO is prone to assign large weights to outliers, resulting in both a drop in performance, and training instability. To provide empirical insights into how outliers affect DRO, in Section 3 we conducted experiments examining how the performance of DRO changes as we removed or added outliers

---

[*]Equal contribution [1]School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Runtian Zhai <rzhai@cmu.edu>.
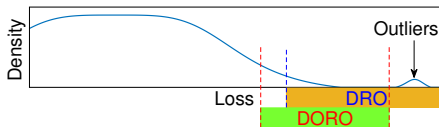
Figure 1. DORO avoids overfitting to outliers.

to the dataset. The results of these experiments indicate that outliers bring about the observed bad performance of DRO. Thus, it is crucial to first enhance the robustness of DRO to outliers before applying it to real-world applications.

To this end, we propose DORO, an outlier robust refinement of DRO which takes inspiration from robust statistics. At the core of this approach is a refined risk function which prevents DRO from overfitting to potential outliers. Intuitively speaking, the new risk function adaptively filters out a small fraction of data with high risk during training, which is potentially caused by outliers. Figure 1 illustrates the difference between DRO and DORO. In Section 4 we implement DORO for the Cressie-Read family of Rényi divergence, and for our theoretical and empirical study we primarily focus on CVaR-DORO and χ^2-DORO. In Section 5 we provide theoretical results guaranteeing that DORO can effectively handle subpopulation shift in the presence of outliers. Then, in Section 6 we empirically demonstrate that DORO improves the performance and stability of DRO. We conduct large-scale experiments on three datasets: the tabular dataset COMPAS, the vision dataset CelebA, and the language dataset CivilComments-Wilds.

Contributions Our contributions are summarized below:

- We demonstrate that the sensitivity of DRO to outliers is a direct cause of the irregular behavior of DRO with some intriguing experimental results in Section 3.

- We propose and implement DORO as an outlier robust refinement of DRO in Section 4. Then, in Section 5 we provide theoretical guarantees for DORO.

- We conduct large-scale experiments in Section 6 and empirically show that DORO improves the performance and stability of DRO. We also analyze the effect of hyperparameters on DRO and DORO.

Related Work Distributional shift naturally arises in many machine learning applications and has been widely studied in statistics, applied probability and optimization (Shimodaira, 2000; Huang et al., 2006; Bickel et al., 2007; Quionero-Candela et al., 2009). One common type of distributional shift is *domain generalization* where the training and testing distributions consist of distinct domains, and relevant topics include domain adaptation (Patel et al., 2015; Wang & Deng, 2018) and transfer learning (Pan & Yang,

2009; Tan et al., 2018). Another common type of distributional shift studied in this paper is subpopulation shift, where the two distributions consist of the same group of domains. Subpopulation shift is closely related to algorithmic fairness and class imbalance. For algorithmic fairness, a number of fairness notions have been proposed, such as individual fairness (Dwork et al., 2012; Zemel et al., 2013), group fairness (Hardt et al., 2016; Zafar et al., 2017), counterfactual fairness (Kusner et al., 2017) and Rawlsian Max-Min fairness (Rawls, 2001; Hashimoto et al., 2018). The setting of subpopulation shift is most closely related to the Rawlsian Max-Min fairness notion. Several recent papers (Hashimoto et al., 2018; Oren et al., 2019; Xu et al., 2020) proposed using DRO to deal with subpopulation shift, but it was also observed that DRO was prone to overfit in practice (Sagawa et al., 2020a;b). (Hashimoto et al., 2018) raised the open question whether it is possible to design algorithms both fair to unknown latent subpopulations and robust to outliers, and this work answers this question positively.

Outlier robust estimation is a classic problem in statistics starting with the pioneering works of (Tukey, 1960; Huber, 1992). Recent works in statistics and machine learning (Lai et al., 2016; Diakonikolas et al., 2017; Prasad et al., 2018; Diakonikolas et al., 2019) provided efficiently computable outlier-robust estimators for high-dimensional mean estimation with corresponding error guarantees. Outliers have a greater effect on the performance of DRO than ERM (Hu et al., 2018), due to its focus on the tail performance, so removing this negative impact of outliers is crucial for the success of DRO in its real-world applications. One closely related recent work is (Lee et al., 2020), and DORO can be viewed as a combination of risk-averse and risk-seeking methods discussed in this paper.

2. Background

This section provides the necessary background of subpopulation shift and DRO.

2.1. Subpopulation Shift

A machine learning task with subpopulation shift requires a model that performs well on the data distribution of each subpopulation. Let the input space be $\mathcal{X}$ and the label space be $\mathcal{Y}$. We are given a training set containing m samples i.i.d. sampled from some data distribution P over $\mathcal{X} \times \mathcal{Y}$. There are K predefined domains (subpopulations) $\mathcal{D}_1, \cdots, \mathcal{D}_K$, each of which is a subset of $\mathcal{X} \times \mathcal{Y}$. For example, in an algorithmic fairness task, domains are demographic groups defined by a number of *protected features* such as race and sex. Let $P_k(z) = P(z|z \in \mathcal{D}_k)$ be the conditional training distribution over $\mathcal{D}_k$, where $z = (x, y)$. The goal is to train a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by $\theta \in \Theta$ that performs well over every P_k. Denote the expected risk

over P by $\mathcal{R}(\theta; P) = \mathbb{E}_{Z \sim P}[\ell(\theta; Z)]$ where $\ell(\theta; z)$ is a measurable loss function. Then the expected risk over P_k is $\mathcal{R}_k(\theta; P) = \mathbb{E}_{Z \sim P_k}[\ell(\theta; Z)]$. The objective is to minimize the *worst-case risk* defined as

$$\mathcal{R}_{\max}(\theta; P) = \max_{k=1,\cdots,K} \mathcal{R}_k(\theta; P) \tag{1}$$

Several different settings were studied by previous work:

Overlapping vs Non-overlapping   The *overlapping* setting allows the domains to overlap with each other while *non-overlapping* does not. For example, suppose we have two protected features: race (White and Others) and sex (Male and Female). Under either setting we will have four domains. Under the overlapping setting we will have *White*, *Others*, *Male* and *Female*, while under the non-overlapping setting we will have *White Male*, *White Female*, *Others Male* and *Others Female*. All the experiments in this work are conducted under the overlapping setting. Each instance may belong to zero, one or more domains.

Domain-Aware vs Domain-Oblivious   Some previous work has assumed that domain memberships of instances are known at least during training. This is called the *domain-aware* setting. However, (Hashimoto et al., 2018) argue that in many real applications, domain memberships are unknown during training, either because it is hard to extract the domain information from the input, or because it is hard to identify all protected features. Thus, a line of recent work (Hashimoto et al., 2018; Lahoti et al., 2020) studies the *domain-oblivious* setting, in which the training algorithm does not know the domain membership of any instance (even the number of domains K is unknown). In this work, we focus on the domain-oblivious setting.

2.2. Distributionally Robust Optimization (DRO)

Under the domain-oblivious setting, we cannot compute the worst-case risk since we have no access to $\mathcal{D}_1, \cdots, \mathcal{D}_K$. In this case, the framework of DRO instead maximizes the performance over the worst-off subpopulation in general. Specifically, given some divergence D between distributions, DRO aims to minimize the expected risk over the worst-case distribution Q (that is absolutely continuous with respect to training distribution P, so that $Q \ll P$) in a ball w.r.t. divergence D around the training distribution P.

Thus, while empirical risk minimization (ERM) algorithm minimizes the expected risk $\mathcal{R}(\theta; P)$, DRO minimizes the *expected DRO risk* defined as:

$$\mathcal{R}_{D,\rho}(\theta; P) = \sup_{Q \ll P} \{\mathbb{E}_Q[\ell(\theta; Z)] : D(Q \parallel P) \leq \rho\} \tag{2}$$

for some $\rho > 0$. Different divergence functions D derive different DRO risks. In this work, we focus on the Cressie-

Read family of Rényi divergence (Cressie & Read, 1984) formulated as:

$$D_\beta(Q \parallel P) = \int f_\beta(\frac{dQ}{dP})dP \tag{3}$$

where $\beta > 1$, and $f_\beta(t)$ is defined as:

$$f_\beta(t) = \frac{1}{\beta(\beta - 1)} \left(t^\beta - \beta t + \beta - 1\right) \tag{4}$$

An advantage of the Cressie-Read family is that it has the following convenient dual characterization (see Lemma 1 of (Duchi & Namkoong, 2018) for the proof):

$$\mathcal{R}_{D_\beta,\rho}(\theta; P) = \inf_{\eta \in \mathbb{R}} \left\{ c_\beta(\rho)\mathbb{E}_P[(l(\theta; Z) - \eta)_+^{\beta_*}]^{\frac{1}{\beta_*}} + \eta \right\} \tag{5}$$

where $\beta_* = \frac{\beta}{\beta - 1}$, and $c_\beta(\rho) = (1 + \beta(\beta - 1)\rho)^{\frac{1}{\beta}}$.

The following proposition shows that DRO can handle subpopulation shift under the domain-oblivious setting. The only information DRO needs during training is α, the ratio between the size of the smallest domain and the size of the population. See the proof in Appendix A.1.

Proposition 1. *Let* $\alpha = \min_{k=1,\cdots,K} P(\mathcal{D}_k)$ *be the minimal group size, and define* $\rho = f_\beta(\frac{1}{\alpha})$*. Then*

$$\mathcal{R}_{\max}(\theta; P) \leq \mathcal{R}_{D_\beta,\rho}(\theta; P) \tag{6}$$

While the Cressie-Read formulation only defines the f-divergence for finite $\beta \in (1, +\infty)$, it can be shown that the dual characterization is valid for $\beta = \infty$ as well, for which the DORO risk becomes the well-known *conditional value-at-risk* (CVaR) (See e.g. (Duchi & Namkoong, 2018), Example 3). In our theoretical analysis and experiments, we delve into two most widely-used sepecial cases of the Cressie-Read family: (i) $\beta = \infty$, which corresponds to CVaR; (ii) $\beta = 2$, which corresponds to χ^2-*DRO risk* used in (Hashimoto et al., 2018). Table 1 summarizes the relevant quatities in these two special cases.

Table 1. CVaR and χ^2-DRO. α is the ratio between the size of the smallest domain and the size of the population.

	CVaR	χ^2-**DRO**
β	∞	2
β_*	1	2
ρ	$-\log(\alpha)$	$\frac{1}{2}(\frac{1}{\alpha} - 1)^2$
$c_\beta(\rho)$	α^{-1}	$\sqrt{1 + (\frac{1}{\alpha} - 1)^2}$
$D_\beta(Q \parallel P)$	$\sup \log \frac{dQ}{dP}$	$\frac{1}{2} \int (dQ/dP - 1)^2 dP$
DRO Risk	$\text{CVaR}_\alpha(\theta; P)$	$\mathcal{R}_{D_{\chi^2},\rho}(\theta; P)$

For example, the dual form of CVaR is

$$\text{CVaR}_\alpha(\theta; P) = \inf_{\eta \in \mathbb{R}} \{\alpha^{-1}\mathbb{E}_P[(\ell(\theta; Z) - \eta)_+] + \eta\} \tag{7}$$

It is easy to see that the optimal η of (7) is the α-quantile of $l(\theta; Z)$ defined as

$$q_\theta(\alpha) = \inf_q \{ P_{Z \sim P}(\ell(\theta; Z) > q) \le \alpha \} \qquad (8)$$

The dual form (7) shows that CVaR in effect minimizes the expected risk on the worst α portion of the training data.

The following corollary of Proposition 1 shows that both $\text{CVaR}_\alpha(\theta; P)$ and $\mathcal{R}_{D_{\chi^2}, \rho}(\theta; P)$ are upper bounds of $\mathcal{R}_{\max}(\theta; P)$, so that minimizing either of them guarantees a small worst-case risk (see the proof in Appendix A.2):

Corollary 2. *Let* $\alpha = \min_{k=1, \cdots, K} P(\mathcal{D}_k)$ *be the minimal group size, and* $\rho = \frac{1}{2}(\frac{1}{\alpha} - 1)^2$. *Then*

$$\mathcal{R}_{\max}(\theta; P) \le \text{CVaR}_\alpha(\theta; P) \le \mathcal{R}_{D_{\chi^2}, \rho}(\theta; P) \quad (9)$$

3. DRO is Sensitive to Outliers

Although the construction of DRO aims to be effective against subpopulation shift as detailed in the previous section, when applied to real tasks DRO is found to have poor and unstable performance. After some examination, we pinpoint one direct cause of this phenomenon: the vulnerablity of DRO to outliers that widely exist in modern datasets. In this section, we will provide some intriguing experimental results to show that:

1. DRO methods have poor and unstable performances.

2. Sensitivity to outliers is a direct cause of DRO's poor performance. To support this argument, we show that DRO becomes good and stable on a "clean" dataset constructed by removing the outliers from the original dataset, and new outliers added to this "clean" dataset compromise DRO's performance and stability.

We conduct experiments on COMPAS (Larson et al., 2016), a recidivism prediction dataset with 5049 training instances (after preprocessing and train-test splitting). We select two features as protected features: race and sex. The two protected features define four overlapping demographic groups: *White*, *Others*, *Male* and *Female*. A two-layer feed-forward neural network with ReLU activations is used as the classification model. We train three models on this dataset with ERM, CVaR and χ^2-DRO. Then we remove the outliers from the training set using the following procedure: We first train a model with ERM, and then remove 200 training instances that incur the highest loss on this model, as outliers are likely to have poorer fit. Then we reinitialize the model, train it on the new training set with ERM, and remove 200 more instances with the highest loss from the new training set. This process is repeated 5 times, so that 1000 training instances are removed and we get a new training set with
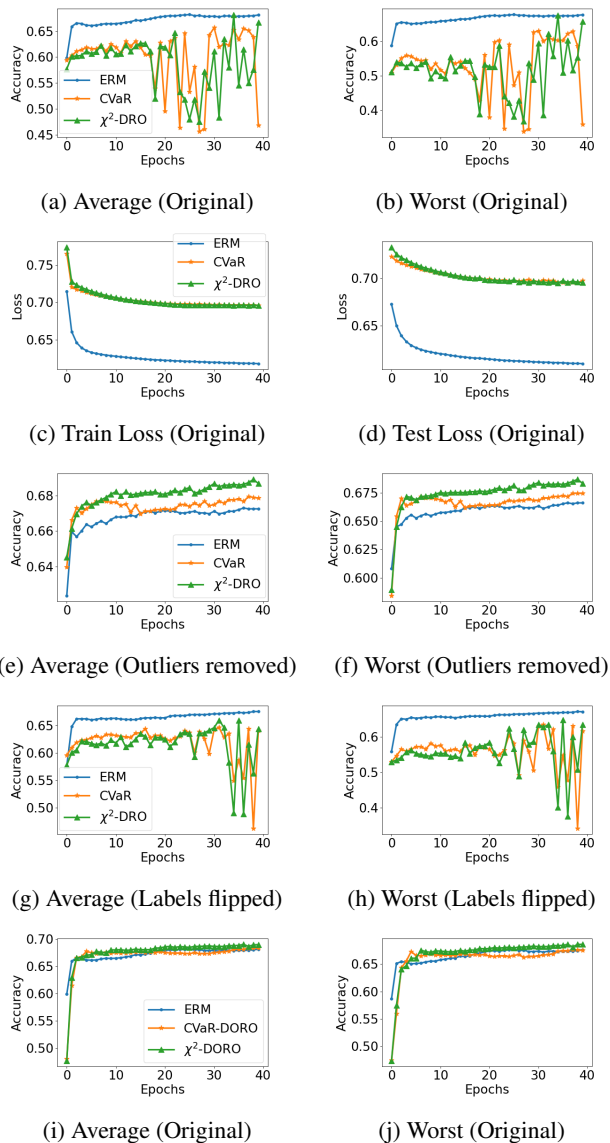

(a) Average (Original) (b) Worst (Original)

(c) Train Loss (Original) (d) Test Loss (Original)

(e) Average (Outliers removed) (f) Worst (Outliers removed)

(g) Average (Labels flipped) (h) Worst (Labels flipped)

(i) Average (Original) (j) Worst (Original)

Figure 2. Average/Worst-case test accuracies on the COMPAS dataset (Original, "clean" with the outliers removed, and "clean with label noise" with 20% of the labels flipped). The second row shows the train/test loss of ERM and DRO on the original dataset (average over all samples). The last row shows the performance of DORO on the original dataset.

4049 instances. Note that this procedure is not guaranteed to remove all outliers and retain all inliers, but is sufficient for the purposes of our demonstration. We then run the three algorithms again on this *same* "clean" training set.

We plot the test accuracies (average and worst across four demographic groups) of the models achieved by the three methods in Figure 2. The first row shows the results on the original dataset, and the second row shows the results on the "clean" dataset with the outliers removed. We can

see that in the first row, for both average and worst-case test accuracies, the DRO curves are below the ERM curves and jumping up and down, which implies that DRO has lower performance than ERM and is very unstable on the original dataset. However, the third row shows that DRO becomes good and stable after the outliers are removed. For comparison, in the second row we plot the train/test loss on the original dataset of the three methods (for ERM we plot the ERM loss, and for DRO we plot the corresponding DRO loss). The train and test losses of DRO descend steadily while the average and worst-case accuracies jump up and down, which indicates that the instability is not an optimization issue, but rather stems from the existence of outliers. It should also be emphasized that these outliers naturally exist in the original dataset since no outliers have been manually added yet.

To further substantiate our conclusion, we consider another common source of outliers: incorrect labels. We randomly flip 20% of the labels of the "clean" COMPAS dataset with the outliers removed, and run the three training methods again. The results are plotted in the fourth row of Figure 2, which shows that while the label noise just slightly influences ERM, it significantly downgrades the performance and stability of the two DRO methods.

Likewise, (Hu et al., 2018) also found in their experiments that DRO had even lower performance than ERM (see their Table 1). Essentially, DRO methods minimize the expected risk on the worst portion of the training data, which contains a higher density of outliers than the whole population. Training on these instances naturally result in the observed bad performance of DRO.

In the next section we will propose DORO as a solution to the problem revealed by the experiments in this section. We plot the performances of the two DORO algorithms we implement in the last row of Figure 2, which compared to the first row shows that DORO improves the performance and stability of DRO on the original dataset.

4. DORO

Problem Setting   The goal is to train a model on a dataset with outliers to achieve high tail performance on the clean underlying data distribution P. Denote the observed contaminated training distribution by P_{train}. We formulate P_{train} with Huber's ϵ-contamination model (Huber, 1992), in which the training instances are i.i.d. sampled from

$$P_{\text{train}} = (1 - \epsilon)P + \epsilon\tilde{P} \qquad (10)$$

where $\tilde{P}$ is an *arbitrary* outlier distribution, and $0 < \epsilon < \frac{1}{2}$ is the noise level. The objective is to minimize $\mathcal{R}_{\max}(\theta; P)$, the worst-case risk over the clean distribution P.

Algorithm 1 DORO with D_β Divergence

Input: Batch size n, outlier fraction ϵ, minimal group size α
for each iteration **do**
    Sample a batch $z_1, \cdots, z_n \sim P_{\text{train}}$
    Compute losses: $\ell_i = \ell(\theta, z_i)$ for $i = 1, \cdots, n$
    Sort the losses: $\ell_{i_1} \geq \cdots \geq \ell_{i_n}$
    Find $\eta^* = \arg\min_\eta F(\theta, \eta)$ where $F(\theta, \eta) = c_\beta(\rho) \cdot [\frac{1}{n-\lfloor\epsilon n\rfloor}\sum_{j=\lfloor\epsilon n\rfloor+1}^n (\ell(\theta; z_{i_j}) - \eta)_+^{\beta_*}]^{1/\beta_*} + \eta$
    Update θ by one step to minimize $\ell(\theta) = F(\theta, \eta^*)$ with some gradient method
end for

DORO Risk   We propose to minimize the following *expected ϵ-DORO risk*:

$$\mathcal{R}_{D,\rho,\epsilon}(\theta; P_{\text{train}}) =$$
$$\inf_{P'}\{\mathcal{R}_{D,\rho}(\theta; P') : \exists \tilde{P}' \text{ s.t. } P_{\text{train}} = (1-\epsilon)P' + \epsilon\tilde{P}'\} \qquad (11)$$

The DORO risk is motivated by the following intuition: we would like the algorithm to avoid the "hardest" instances that are likely to be outliers, and the optimal P' of (11) consists of the "easiest" $(1 - \epsilon)$-portion of the training set given the current model parameters θ. The ϵ in DORO is a hyperparameter selected by the user since the real noise level of the dataset is unknown. Let the real noise level of P_{train} be ϵ_0. For any $\epsilon \geq \epsilon_0$, there exist $\tilde{P}_0$ and $\tilde{P}$ such that $P_{\text{train}} = (1 - \epsilon_0)P + \epsilon_0\tilde{P}_0 = (1 - \epsilon)P + \epsilon\tilde{P}$, so we only need to make sure that ϵ is not less than the real noise level.

The following proposition provides the formula for computing the DORO risk for the Cressie-Read family (See the proof in Appendix A.3.1):

Proposition 3. *Let ℓ be a continuous non-negative loss function, and suppose P_{train} is a continuous distribution. Then the formula for computing the DORO risk with D_β is*

$$\mathcal{R}_{D_\beta,\rho,\epsilon}(\theta; P_{\text{train}}) =$$
$$\inf_\eta\{c_\beta(\rho)\mathbb{E}_{Z\sim P_{\text{train}}}[(\ell(\theta; Z) - \eta)_+^{\beta_*} \mid \qquad (12)$$
$$P_{Z'\sim P_{\text{train}}}(\ell(\theta; Z') > \ell(\theta; Z)) \geq \epsilon]^{\frac{1}{\beta_*}} + \eta\}$$

Remark   In Proposition 3, we assume the continuity of P_{train} to keep the formula simple. For an arbitrary distribution P_{train}, we can obtain a similar formula, but the formula is much more complex than (12). The general formula can be found in Appendix A.3.2.

With this formula, we develop Algorithm 1. In the algorithm, we first order the batch samples according to their training losses, then find the optimal η^* using some numerical method (we use Brent's method (Brent, 1971) in our implementation), and finally update θ with some gradient method.

Note that generally it is difficult to find the minimizer of the DORO risk for neural networks, and our algorithm is inspired by the ITLM algorithm (Shen & Sanghavi, 2019), in which they proved that the optimization converges to ground truth for a few simple problems. Particularly, using the quantities listed in Table 1, we can implement CVaR-DORO and χ^2-DORO. In the sections that follow, we will focus on the performances of CVaR-DORO and χ^2-DORO in particular. We denote the CVaR-DORO risk by $\text{CVaR}_{\alpha,\epsilon}(\theta; P_{\text{train}})$, and the χ^2-DORO risk by $\mathcal{R}_{D_{\chi^2},\rho,\epsilon}(\theta; P_{\text{train}})$.

5. Theoretical Analysis

Having the DORO algorithms implemented, in this section we prove that DORO can effectively handle subpopulation shift in the presence of outliers. The proofs to the results in this section can be found in Appendix A.4. We summarize our theoretical results as follows:

1. The minimizer of DORO over the contaminated distribution P_{train} achieves a DRO risk close to the minimum over the clean distribution P (Theorem 5). We complement our analysis with information-theoretical lower bounds (Theorem 6) implying that the optimality gaps given by Theorem 5 are optimal.

2. The worst-case risk $\mathcal{R}_{\max}$ over P is upper bounded by the DORO risk over P_{train} times a constant factor (Theorem 7). This result parallels Corollary 2 in the uncontaminated setting and guarantees that minimizing the DORO risk over P_{train} effectively minimizes $\mathcal{R}_{\max}$ over P.

Our results are based on the following lemma which lower bounds the DORO risk over P_{train} by the infimum of the original DRO risk in a TV-ball centered at P:

Lemma 4. *Let* $\text{TV}(P, Q) = \frac{1}{2} \int_{\mathcal{X} \times \mathcal{Y}} |P(z) - Q(z)| dz$ *be the total variation, and* P_{train} *be defined by (10). Then the DORO risk can be lower bounded by:*

$$\mathcal{R}_{D,\rho,\epsilon}(\theta; P_{\text{train}}) \geq$$
$$\inf_{P''}\{\mathcal{R}_{D,\rho}(\theta; P'') : \text{TV}(P, P'') \leq \frac{\epsilon}{1-\epsilon}\} \quad (13)$$

The main results we are about to present only require very mild assumptions. For the first result, we assume that ℓ has a bounded $(2k)$-th moment on P, a standard assumption in the robust statistics literature:

Theorem 5. *Let* P_{train} *be defined by (10). Denote the minimizer of the DORO risk by* $\hat{\theta}$. *If* ℓ *is non-negative, and* $\ell(\hat{\theta}; Z)$ *has a bounded* $(2k)$*-th moment:* $\mathbb{E}_{Z \sim P}[l(\hat{\theta}; Z)^{2k}] = \sigma_{2k}^{2k} < +\infty$, *then we have:*

$$\text{CVaR}_\alpha(\hat{\theta}; P) - \inf_\theta \text{CVaR}_\alpha(\theta; P) \leq O_{\alpha,k}(1)\sigma_{2k}\epsilon^{1-\frac{1}{2k}} \quad (14)$$

and if $k > 1$, *then we have:*

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta}; P) - \inf_\theta \mathcal{R}_{D_{\chi^2},\rho}(\theta; P) \leq O_{\rho,k}(1)\sigma_{2k}\epsilon^{\left(\frac{1}{2} - \frac{1}{2k}\right)} \quad (15)$$

Furthermore, the above optimality gaps are optimal:

Theorem 6. *There exists a pair of* (P, P_{train}) *where* $P_{\text{train}} = (1-\epsilon)P + \epsilon P'$ *and* P *has uniformly bounded* $2k$*-th moment:* $\forall \theta \in \Theta$, $\mathbb{E}_P[l(\theta, Z)^{2k}] \leq \sigma_{2k}^{2k}$ *such that for any learner with only access to* P_{train}, *the best achievable error in DRO over* P *is lower bounded by*

$$\text{CVaR}_\alpha(\hat{\theta}; P) - \inf_{\theta \in \Theta} \text{CVaR}_\alpha(\theta; P) \geq \Omega_{\alpha,k}(1)\sigma_{2k}\epsilon^{1-\frac{1}{2k}} \quad (16)$$

$$\mathcal{R}_{D_{\chi^2},\rho}(\hat{\theta}; P) - \inf_{\theta \in \Theta} \mathcal{R}_{D_{\chi^2},\rho}(\theta; P) \geq \Omega_{\rho,k}(1)\sigma_{2k}\epsilon^{\left(\frac{1}{2} - \frac{1}{2k}\right)} \quad (17)$$

We make a few remarks on these theoretical results. The $O(\epsilon^{1-\frac{1}{2k}})$ and $O(\epsilon^{\frac{1}{2}-\frac{1}{2k}})$ rates resemble the existing works on robust mean/moment estimation, see e.g. (Kothari et al., 2018; Prasad et al., 2020). The robust mean estimation problem can be seen as a special case of CVaR when $\alpha = 1$, where CVaR of any θ is just the mean of $l(\theta, Z)$. On the other hand, the connection between CVaR and robust moment estimation can be built with the dual characterization (5): for any *fixed* dual variable η, evaluating the dual is nothing but a robust (β_*-th) moment estimation of the random variable $(l(\theta, Z) - \eta)_+$. However, the problem we are trying to tackle in the above theorems is more challenging, in the sense that (1) DRO risk involves taking infimum over all $\eta \in \mathbb{R}$, but the moments of $(l(\theta, Z) - \eta)_+$ are not *uniformly* bounded for all possible η's; and (2) the optimal dual variable η^* can be very different even for distributions extremely close in total-variation distance. In Appendix A.4 we discuss how to overcome these difficulties in detail.

Our second result is a robust analogue to Corollary 2: we show that the worst-case risk $\mathcal{R}_{\max}$ can be upper bounded by a constant factor times the DORO risk $\text{CVaR}_{\alpha,\epsilon}$, under the very mild assumption that ℓ has a uniformly bounded second moment on P and $\mathcal{R}_{\max}$ is not exceedingly small:

Theorem 7. *Let* P_{train} *be defined by (10). Let* $\alpha = \min_{k=1,\cdots,K} P(\mathcal{D}_k)$, *and* $\rho = \frac{1}{2}(\frac{1}{\alpha} - 1)^2$. *If* $\ell(\theta; Z)$ *is a non-negative loss function with a uniformly bounded second moment:* $\mathbb{E}_{Z \sim P}[\ell(\theta; Z)^2] \leq \sigma^2$ *for all* θ, *then we have:*

$$\mathcal{R}_{\max}(\theta; P) \leq \max\{3\text{CVaR}_{\alpha,\epsilon}(\theta; P_{\text{train}}), 3\alpha^{-1}\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}\}$$

$$\leq \max\{3D_{\chi^2,\rho,\epsilon}(\theta; P_{\text{train}}), 3\alpha^{-1}\sigma\sqrt{\frac{\epsilon}{1-\epsilon}}\} \quad (18)$$

Note that a similar result can be derived under the bounded $2k$-th moment condition with different constants.

6. Experiments

In this section, we conduct large-scale experiments on modern datasets. Our results show that DORO improves the performance and stability of DRO. We also analyze the effect of hyperparameters on DRO and DORO.

6.1. Setup

Datasets   Our goal is to apply DRO to real tasks with subpopulation shift on modern datasets. While many previous work used small tabular datasets such as COMPAS, these datasets are insufficient for our purpose. Therefore, apart from COMPAS, we use two large datasets: CelebA (Liu et al., 2015) and CivilComments-Wilds (Borkan et al., 2019; Koh et al., 2020). CelebA is a widely used vision dataset with 162,770 training instances, and CivilComments-Wilds is a recently released language dataset with 269,038 training instances. Both datasets are captured in the wild and labeled by potentially biased humans, so they can reveal many challenges we need to face in practice.

We summarize the datasets we use as follows: (i) COMPAS: recidivism prediction, where the target is whether the person will reoffend in two years; (ii) CelebA: human face recognition, where the target is whether the person has blond hair; (iii) CivilComments-Wilds: toxicity identification, where the target is whether the user comment contains toxic contents. All targets are binary. For COMPAS, we randomly sample 70% of the instances to be the training data (with a fixed random seed) and the rest is the validation/testing data. Both CelebA and CivilComments-Wilds have official train-validation-test splits, so we use them directly.

Domain Definition   On COMPAS we define 4 domains (subpopulations), and on CelebA and CivilComments-Wilds we define 16 domains for each. Our domain definitions cover several types of subpopulation shift, such as different demographic groups, class imbalance, labeling biases, confounding variables, etc. See Appendix B.1 for details.

Training   We use a two-layer feed-forward neural network activated by ReLU on COMPAS, a ResNet18 (He et al., 2016) on CelebA, and a BERT-base-uncased model (Devlin et al., 2019) on CivilComments-Wilds. On each dataset, we run ERM, CVaR, χ^2-DRO, CVaR-DORO and χ^2-DORO. Each algorithm is run 300 epochs on COMPAS, 30 epochs on CelebA and 5 epochs on CivilComments-Wilds. For each method we collect the model achieved at the end of every epoch, and select the best model through validation. (On CivilComments-Wilds we collect 5 models each epoch, one for every ~20% of the training instances.)

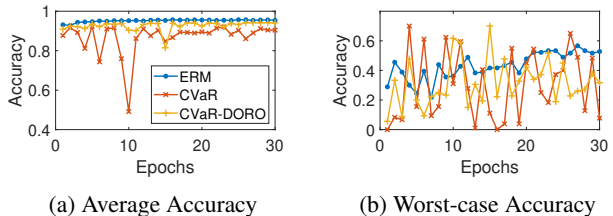**Model Selection**   To select the best model, we assume that the domain membership of each instance is available

(a) Average Accuracy   (b) Worst-case Accuracy

Figure 3. Test accuracies of CVaR and CVaR-DORO on CelebA ($\alpha = 0.1$, $\epsilon = 0.01$).

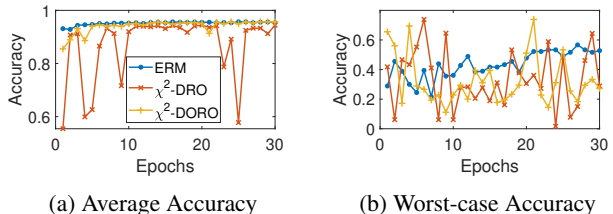(a) Average Accuracy   (b) Worst-case Accuracy

Figure 4. Test accuracies of χ^2-DRO and χ^2-DORO on CelebA ($\alpha = 0.3$, $\epsilon = 0.01$).

in the validation set, and select the model with the highest worst-case validation accuracy. This is an oracle strategy since it requires a domain-aware validation set. Over the course of our experiments, we have realized that model selection with no group labels during validation is a very hard problem. On the other hand, model selection has a huge impact on the performance of the final model. We include some preliminary discussions on this issue in Appendix B.2. Since model selection is not the main focus of this paper, we pose it as an open question.

6.2. Results

The 95% confidence intervals of the mean test accuracies on each dataset are reported in Table 2. For every DRO and DORO method, we do a grid search to pick the best α and ϵ that achieve the best worst-case accuracy (see the optimal hyperparameters in Appendix B.3). Each experiment is repeated 10 times on COMPAS and CelebA, and 5 times on CivilComments-Wilds with different random seeds. Table 2 clearly shows that on all datasets, DORO consistently improves the average and worst-case accuracies of DRO.

Next, we analyze the stability of the algorithms on the CelebA dataset. We use the α that achieves the optimal DRO performance for each of CVaR and χ^2-DRO, and compare them to DORO with the same value of α and $\epsilon = 0.01$. χ^2-DRO achieves its optimal performance with a bigger α than CVaR because it is less stable. To quantitatively compare the stability, we compute the standard deviations of the test accuracies across epochs and report the results in Table 3. To further visualize the training dynamics, we run all algorithms with one fixed random seed, and plot the test accuracies during training in Figures 3 and 4. Table 3 shows that the standard deviation of the test accuracy of DORO is smaller and in Figures 3a and 4a the DORO curves

Table 2. The average and worst-case test accuracies of the best models achieved by different methods. (%)

Dataset	Method	Average Accuracy	Worst-case Accuracy
COMPAS	ERM	69.31 ± 0.19	68.83 ± 0.18
	CVaR	68.52 ± 0.31	68.22 ± 0.30
	CVaR-DORO	69.38 ± 0.10	69.11 ± 0.05
	χ^2-DRO	67.93 ± 0.40	67.32 ± 0.60
	χ^2-DORO	69.62 ± 0.16	69.22 ± 0.11
CelebA	ERM	95.01 ± 0.38	53.94 ± 2.02
	CVaR	82.83 ± 1.33	66.44 ± 2.34
	CVaR-DORO	92.91 ± 0.48	72.17 ± 3.14
	χ^2-DRO	83.85 ± 1.42	67.76 ± 3.22
	χ^2-DORO	82.18 ± 1.17	68.33 ± 1.79
CivilComments-Wilds	ERM	92.04 ± 0.24	64.62 ± 2.48
	CVaR	89.11 ± 0.76	63.90 ± 4.42
	CVaR-DORO	90.45 ± 0.70	68.00 ± 2.10
	χ^2-DRO	90.08 ± 0.92	65.55 ± 1.51
	χ^2-DORO	90.11 ± 1.09	67.19 ± 2.51

Table 3. Standard deviations of average/worst-case test accuracies during training on CelebA. ($\alpha = 0.1$ for CVaR/CVaR-DORO; $\alpha = 0.3$ for χ^2-DRO/χ^2-DORO. $\epsilon = 0.01$) (%)

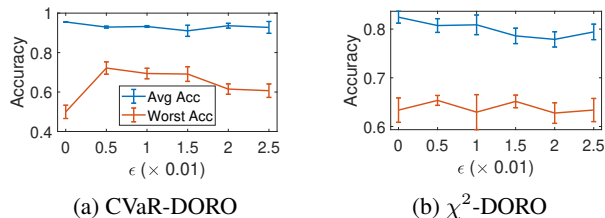| Method	Average	Worst-case
ERM	0.73 ± 0.06	8.59 ± 0.90
CVaR	11.53 ± 1.72	21.47 ± 0.71
CVaR-DORO	4.03 ± 1.57	16.84 ± 0.91
χ^2-DRO	8.88 ± 2.98	19.06 ± 1.18
χ^2-DORO	1.60 ± 0.34	13.01 ± 1.40

(a) CVaR-DORO   (b) χ^2-DORO

Figure 5. Effect of ϵ on the test accuracies of CVaR/χ^2-DORO on CelebA ($\alpha = 0.2$). DORO with $\epsilon = 0$ is equivalent to DRO.

(a) CVaR   (b) χ^2-DRO

(c) CVaR-DORO   (d) χ^2-DORO

Figure 6. Effect of α on the test accuracies of DRO and DORO on CelebA ($\epsilon = 0.01$).

are flatter than the DRO curves, which implies that DORO improves the stability of DRO. Although it is hard to tell whether DORO has a more stable worst-case accuracy from the figures, our quantitative results in Table 3 confirm that DORO has more stable worst-case test accuracies.

6.3. Effect of Hyperparameters

In this part, we study how α and ϵ affect the test accuracies of DORO with two experiments on CelebA, providing insight into how to select the optimal hyperparameters.

In the first experiment, we fix $\alpha = 0.2$, and run the two DORO algorithms with different values of ϵ. The results are plotted in Figure 5. We can see that for both methods, as ϵ increases, the average accuracy slightly decreases, while

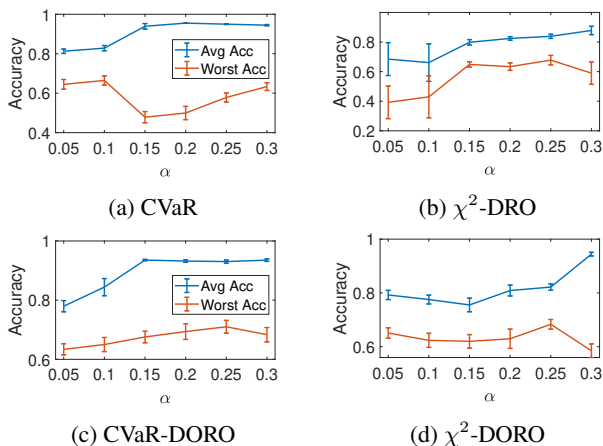the worst-case accuracy first rises and then drops. Both average and worst-case accuracies will drop if ϵ is too big. Moreover, both methods achieve the optimal worst-case accuracy at $\epsilon = 0.005$. We conjecture that the real noise level of the CelebA dataset is around 0.005, and that the optimal ϵ should be close to the real noise level.

In the second experiment, we run DRO and DORO ($\epsilon = 0.01$) with different values of α. The results are plotted in Figure 6. First, we observe that for all methods, the optimal α is much bigger than the real α of the dataset. The real α of the CelebA dataset is around 0.008 (see Appendix B.1, Table 1), much smaller than those achieving the highest worst-case accuracies in the figures. Second, in all four figures the overall trend of the average accuracy is that it grows with α. Third, both CVaR-DORO and χ^2-DORO achieve the optimal worst-case accuracy at $\alpha = 0.25$, but the worst-case accuracy drops as α goes to 0.3.

7. Discussion

In this work we pinpointed one direct cause of the performance drop and instability of DRO: the sensitivity of DRO to outliers in the dataset. We proposed DORO as an outlier robust refinement of DRO, and implemented DORO for the Cressie-Read family of Rényi divergence. We made a positive response to the open question raised by (Hashimoto et al., 2018) by demonstrating the effectiveness of DORO both theoretically and empirically.

One alternative approach to making DRO robust to outliers is removing the outliers from the dataset via preprocessing. In Section 3 we used a simple version of iterative trimming (Shen & Sanghavi, 2019) to remove outliers from the training set. Compared to iterative trimming, DORO does not require retraining the model and does not throw away any data. In addition, preprocessing methods such as iterative trimming cannot cope with online data (where new instances are received sequentially), but DORO is still feasible.

The high-level idea of DORO can be extended to other algorithms that deal with subpopulation shift, such as static reweighting (Shimodaira, 2000), adversarial reweighting (Hu et al., 2018; Lahoti et al., 2020) and group DRO (Sagawa et al., 2020a). The implementations might be different, but the basic ideas are the same: to prevent the algorithm from overfitting to potential outliers. We leave the design of such algorithms to future work.

There is one large open question from this work. In our experiments, we found that model selection without domain information in the validation set is very hard. In Appendix B.2 we study several strategies, such as selecting the model with the lowest CVaR risk or the lowest CVaR-DORO risk, but none of them is satisfactory. A recent paper (Michel et al., 2021) proposed two selection methods Minmax and Greedy-Minmax, but their performances are still much lower than the oracle's (see their Table 2a). (Gulrajani & Lopez-Paz, 2021) also pointed out the difficulty of model selection in domain-oblivious distributional shift tasks. Thus, we believe this question to be fairly non-trivial.

Acknowledgements

References

Barocas, S. and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pp. 81–88, 2007.

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500, 2019.

Brent, R. P. An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14 (4):422–425, 1971.

Cressie, N. and Read, T. R. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464, 1984.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 999–1008, International Convention Centre, Sydney, Australia, 06–11 Aug 2017.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Duchi, J. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4): 463–484, 2011.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 3315–3323. Curran Associates, Inc., 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In Dy, J. and Krause, A. (eds.), *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pp. 492–518. Springer, 1992.

Japkowicz, N. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56. Citeseer, 2000.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Beery, S., et al. Wilds: A benchmark of in-the-wild distribution shifts. *arXiv preprint arXiv:2012.07421*, 2020.

Kothari, P. K., Steinhardt, J., and Steurer, D. Robust moment estimation and improved clustering via sum of squares. In Diakonikolas, I., Kempe, D., and Henzinger, M. (eds.), *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*, pp. 1035–1046. ACM, 2018.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in neural information processing systems*, pp. 4066–4076, 2017.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Lai, K. A., Rao, A. B., and Vempala, S. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 665–674. IEEE, 2016.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

Lee, J., Park, S., and Shin, J. Learning bounds for risk-sensitive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13867–13879. Curran Associates, Inc., 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Michel, P., Hashimoto, T., and Neubig, G. Modeling the second player in distributionally robust optimization. In *International Conference on Learning Representations*, 2021.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in neural information processing systems*, pp. 2208–2216, 2016.

Oren, Y., Sagawa, S., Hashimoto, T., and Liang, P. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4227–4237, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1432.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Prasad, A., Balakrishnan, S., and Ravikumar, P. A robust univariate mean estimator is all you need. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4034–4044. PMLR, 26–28 Aug 2020.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Rawls, J. *Justice as fairness: A restatement.* Harvard University Press, 2001.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020a.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 13–18 Jul 2020b.

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pp. 5739–5748. PMLR, 2019.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. A survey on deep transfer learning. In *International conference on artificial neural networks*, pp. 270–279. Springer, 2018.

Tukey, J. W. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pp. 448–485, 1960.

Wang, M. and Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

Xu, Z., Dan, C., Khim, J., and Ravikumar, P. Class-weighted classification: Trade-offs and robust approaches. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10544–10554. PMLR, 13–18 Jul 2020.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Zhu, B., Jiao, J., and Steinhardt, J. Generalized resilience and robust statistics, 2020.